A Work Project, presented as part of the requirements for the Award of a Master Degree in Finance from the NOVA School of Business And Economics

CREDIT RISK MODELING:
PREDICITING CUSTOMER LOAN DEFAULTS WITH MACHINE LEARNING MODELS

Directed Research Format Nova SBE & zeb consulting

Thomas DORNIGG, 41727

A project carried out on the Master in Finance Program, under the supervision of:

zeb.rolfes.schierenbeck.associates gmbh Advisor

Stephan Trempler – Senior Manager Practice Group Risk & Advanced Methodology

Faculty Advisor

Professor João Pedro PEREIRA

December 15, 2021

Abstract

The assessment of financial credit risk constitutes an important, yet challenging research topic across multiple disciplines. This paper evaluates the risk of customers not being able to repay their obligation on time by utilizing a variety of both parametric and non-parametric (supervised) machine learning models. These methods include Decision Tree, Random Forest, AdaBoost, XGBoost, and Support Vector Machine. In addition, as a benchmark classifier, the traditional credit-risk method, Logistic Regression, was used to perform a comparison. Random Forest and XGBoost outperformed the other methods constantly, provided that thorough data analysis, pre-processing, and model-training are performed.

Keywords: credit risk prediction, credit default, credit scoring, supervised machine learning, binary classification, model validation, graphical user interface

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209

Chapter 1

Introduction

In this chapter, an overview of the main purpose of the thesis is provided. Additionally, topics such as the thesis' background, set-up, and scope will be addressed briefly.

1.1 Background

In today's economic and business environment, financial institutions play a crucial role, most notably when it comes to the decision of allocating financial resources to customers. Within this domain, lending and borrowing are the two key pillars of any well-functioning banking system. A major issue that financial enterprises have to face in this process, once a loan is granted, is the risk that the counterparty is unable to honor its obligation on time. There can be many reasons for a default - in most cases though, the obligor is in a financially stressed situation and may has to face a bankruptcy procedure (Gerstel and Baesens, 2009). Therefore, a significant component of a bank's risk lies in the quality of its assets that needs to be in line with the bank's risk appetite (Hasan, 2016). Hence, a good credit risk management system is a crucial part of any consumer lending process since it attempts to predict the repayment probability of its borrowers before a loan agreement is achieved (Kennedy et al., 2013).

Especially since the subprime mortgage crisis of 2008, credit and default risk have been at the forefront of the worldwide financial news. Regulators and governments realized that one of the main causes of the crisis was due to an overflow of subprime mortgages which were granted to lenders with no substantial financial cushion - later known under the term *NINJA-credits* ("no income, no job and assets"). Soon after the crisis was managed, the spillovers to the real

economy were made evident by a decrease in lending activities (credit crunch) and high levels of non-performing loans (NPLs) on the balance sheets, particularly from European banks. Thus, in order to prevent similar crises in the future and to restore trust in the financial sector, policy-makers strengthened the framework, which, among others, determines the minimum capital requirement banks must hold to cover potential future losses. With the implementation of Basel II, which provides a general framework for supervisory standards and risk management techniques as a guideline for banks to manage and quantify their risks, two options for financial institutions were proposed to realistically assess their regulatory risk exposure: The first is the standardised approach, and the second the internal rating-based (IRB) approach (Granström and Abrahamsson, 2019). Depending on the applied IRB model, foundation or advanced IRB, banks were allowed to compute their own probability of default (PD), loss given default (LGD), and exposure-at-default (EAD) levels. Having said this, only banks meeting certain minimum conditions, such as disclosure requirements and approval from their national supervisors, are permitted to utilize this approach (Gavalas and Syriopoulos, 2014). As summarized by Qader and Sihver (2020), an overview of the models is provided in Table B1.

While the IRB system is not the only admissible procedure for institutions to calculate regulatory capital requirements, it is, however, one of the most frequently used approaches, especially in its original form, where only the PD is an input to be estimated by the institution (Alonso and Carbó, 2020). As Kennedy et al. (2013) highlight, modeling probability of defaults is essentially a discrimination problem (good or bad), consequently one may resort to the numerous classification techniques that have been suggested in the literature.

Given that the credit exposure of financial institutions often amounts to several billions of dollars, even a marginal improvement of a fraction of percentage point in accuracy or precision may resort to a significant decrease in potential future losses and to a decrease in a banks' funding costs (Rochet, 1992). This has motivated both researchers and practitioners to investigate further scoring models to improve accuracy in their risk models (Chopra and Bhilare, 2018). For instance, Steenackers and Goovaerts (1989) suggest a credit scoring model to estimate the probability of default for personal loans which is based on logistic regression. Historically, multivariate discriminant analysis or logistic regressions, such as Probit and/or Logit, have been indeed one of the most commonly used quantitative estimation approaches under the IRB system. Therefore, the goal of this thesis is to investigate more advanced credit risk models, not only their statistical but also technical characteristics, and to evaluate, if they can challenge traditionally applied techniques.

1.2 Purpose

The main objective of this thesis is to investigate which supervised machine learning classifier(s), from a given set of models, perform(s) best in predicting customer loan defaults. Thus, this thesis contributes to the literature by addressing the following two research questions:

RQ1: If machine learning based classifiers are better able to predict applicants' loan defaults than traditional empirical models, how well do they perform in their assessment?

RQ2: Which credit variables/features are the key drivers for customer loan defaults?

1.3 Set-up and scope of this thesis

The scope of this thesis is to train a set of binary (supervised) machine learning classifiers and evaluate the performance regarding their predictive power to classify potential customer loan-defaulters. The novelty of this approach lies in the combination of data preprocessing algorithms via pipelines, which will deal both with the feature engineering and feature selection part. Additionally, great attention is given to model interpretability and feature explainability by utilizing state-of-the-art global and local interpretation techniques such as Shapley Additive Values. A detailed explanation of the used algorithms, as well as the evaluation-metrics, will be given in later sections of this paper.

The remainder of this thesis is organized as follows: section 2 provides a brief literature review on credit risk and previously conducted work in this field. In section 3, the machine learning classifiers, as well as traditional scoring models, will be introduced. Also, a brief introduction to classification evaluation metrics is provided. In section 4, a description of the dataset in combination with an Exploratory Data Analysis (EDA) of the key variables will be presented. In addition, the chapter also deals with the data processing of the main variables such as handling of missing values, feature scaling, outlier-removal, feature selection, and dummy-encoding. Moreover, the handling of imbalanced data via SMOTE and its potential impact on the final result is briefly addressed. Later on, in section 5, the findings of the previous chapter will be discussed and evaluated. Additionally, based on the results of the best classification models, an application of credit scoring via a Graphical User Interface (GUI) will be shown. Finally, in the last part of this paper, section 6, a comprehensive outlook and discussion will be presented.

Chapter 2

Literature Review

This section provides the reader with necessary background information regarding a general definition of credit risk as well as a historic overview of the development of risk models.

2.1 Definition of credit risk (management)

Financial institutions and especially banks face various elements of risk that require to be identified, understood, and carefully managed. According to the Basel II Capital Accord (BCBS, 2006), three main sources of risk can be identified: credit risk, market risk, and operational risk. In terms of potential future losses, *credit risk* shines out as by far the most important one. The Basel Committee on Banking Supervision (BCBS, 1999) defines credit risk as "*the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms*". However, a general quantification and definition of *credit risk assessment* is not as straightforward as the former. On the one hand, by looking from a purely financial point of view, the main goal of credit risk management is to mitigate or minimize losses and to maximize a bank's risk-adjusted rate of return by maintaining credit risk exposure within acceptable parameters (BCBS, 2000). On the other hand, the main objective slightly differs when looked upon from a data mining or econometrics point of view: As Kennedy et al. (2013) describe, the main goal is to assign loan applicants to one of the two groups - *good* or *bad*. A member of the good group is regarded as a safe borrower who is capable of repaying their financial obligation, while a member of the bad group is considered as a risky borrower who is unlikely to pay pack their obligation on time.

Thus, the goal of financial credit risk management, as Ribeiro et al. (2019) puts it, can be summarized as follows: given a number of companies or customers labeled as bad/good credit or bankrupt/healthy, and a set of financial variables that describe the situation of a company/-customer over a given period, predict the probability that the company/customer may belong to a high-risk group or become bankrupt during the subsequent time period. Generally, this type of modeling problem, where only two categories (0 = good creditor, 1 = bad creditor) can be identified, is referred to in the literature as ***binary classification*** problem.

An important point, which must not be neglected when talking about credit risk modeling is the criteria used to determine when a client is in default. For that, the Basel Committee on Banking Supervision has published two widely accepted definitions, which, across several countries, including the EU (see CRR Article 178 No 575 / 2013), harmonises the way banks need to apply regulatory requirements to their capital positions (EBA, 2016). According to the BCBS (2006) working paper, an exposure should be considered as defaulted when either or both of the following two events are met:

- The financial institution considers that the obligor is unlikely to pay its credit obligations in full, without recourse by the institution to actions such as realising security.
- The obligor is past due more than 90 days on any material credit obligation to the institution.

2.2 Development of credit risk methods

Given the sheer amount of prediction models in the last couple of years, there is a demand for reviewing the latest research conducted in the field of credit risk management. Thus, the literature of previously used techniques and today's commonly utilized methods in risk modeling are investigated in this subsection.

2.2.1 Historical development

One of the earliest research in the field of credit risk assessment can be traced back to FitzPatrick (1932), who tried to predict insolvency based on indicators derived from financial statements of companies. A few years later, the parametric technique of *Linear Discriminant Analysis* (LDA) was first proposed by Fisher (1936). It constitutes a method to discriminate between the two groups of applicants, ideally applicable in the area of credit scoring. Soon after, Altman (1968) published an extension of the discriminant analysis by introducing his well-known z-score models

which measure the probability of any organization going bankrupt by using balance sheet items and corporate income as the main equation inputs. Since the publication of Altman's pioneering paper in 1968, a growing number of various statistical classification and prediction techniques were proposed, which quickly replaced purely judgemental-based credit-granting decisions. For instance, Dumitrescu et al. (2021) mention that around the same time, the use of econometric models gained widespread popularity, when the credit card business arose and an automatised decision process was required. Later, Ohlson (1980) introduced a probabilistic approach to predict the credit-worthiness of companies by incorporating *Logit* or *Multiple Logistic Regression*. Slowly but gradually, logistic regression and linear discriminant analysis, among a few others, became the most commonly used credit scoring models in the financial industry because of their distinctive features which can be found in greater detail in Thomas (2009).

As Donga et al. (2010) further explain, the advantageousness of Logit-models has twofold reasons: robustness and transparency. While other models may yield better prediction outcomes, they are vulnerable to changes in population characteristics. More importantly, though, regulation authorities and supervisors require banks to justify decisions made by their used credit risk method, which is hardly manageable when more advanced algorithms like *artificial neural network* or *support vector machine* are used since those are considered to be so-called **black-box models**. Additionally, given the categorical nature of credit history data and the fact that the covariance matrices of good and bad credit classes are not likely to be equal, the appropriateness of LDA for credit risk modeling has been questioned (Kaya et al., 2008). Furthermore, as Reichert et al. (1983) state, credit risk data is usually not normally distributed, even though this might not be a critical restriction in the actual deployment of such models. However, given those limitations, as Dumitrescu et al. (2021) further highlight, most international banks are still utilizing logistic regression, especially for regulatory scores which estimate the probability of default for capital requirements (Basel III) or for point-in-time estimates of expected credit losses (IFRS 9).

Nevertheless, as seen in section 5, state-of-the-art artificial intelligence methods could yield advantages in terms of predictive gains with respect to logistic and/or linear discriminant function models at the cost of being more complex. As Alonso and Carbó (2020) from the Banco de España elaborate, *this innovation might have a significant impact on the financial industry, even at macro level, as its market-level adoption would determine the calculation of risk-weighted assets (RWA) and their variability.*

2.2.2 Credit risk assessment methods today

With the emergence of Big Data technologies, coupled with fierce competition from financial technology startups, credit scoring was one of the first fields in economics where machine learning techniques were rolled out on a broad scale.

Today, a large variety of different approaches, ranging from (1) heuristic methods to (2) statistical models, (3) causal models or a combination of (1) and (2) or (3) have been employed by financial firms to evaluate the creditworthiness of their borrowers (see Figure A.1). There are many studies on credit risk modeling aiming to investigate the advantageousness of those methods over traditional scoring methodologies and to analyze the underlying causes for them. A selected overview of previous conducted work, as summarised by Tawfik (2019), Chen et al. (2016) and Moradi and Rafiei (2019) is shown in Table B2.

Numerous studies have shown that well-configured and trained models can considerably improve customer loan default predictions, without relying too much on restrictive assumptions (Chen et al., 2016). Some examples, among others are: non-parametric models, such as k-nearest neighbor (Hand and Henley, 1997), decision trees as discussed in (Srinivasan and Kim, 1987), (Satchidananda and Simha, 2006), support vector machines (Kaya et al., 2008) and artificial neural networks (Atiya, 2001), (Li et al., 2002), (Yobas et al., 2000). The overwhelming evidence found by Finlay (2011), Paleologo et al. (2010) and Lessmann et al. (2015) have shown that, in the quest for higher scoring accuracy, the performance of machine learning-based credit risk scoring models has markedly been improved since adopting ensemble (aggregation) methods, particularly bagging and boosting algorithms. Ensemble learners (also referred to as multiple classifier systems) as Song et al. (2020) state, are a set of individually trained classifiers, called base learners, whose decisions are combined to generate an improved outcome, usually by weighted or unweighted voting. However, a general conclusion can be drawn that no algorithm from all the mentioned ones consistently outperforms across different data sets (Chen et al., 2016).

Chapter 3

Predictive Modeling Techniques and Model Evaluation

In this section, relevant background information behind traditional as well as modern classification techniques, their potential drawbacks, and evaluation methods are explained.

3.1 Introduction to Machine Learning

Commonly, machine learning is regarded as an application of artificial intelligence (AI) or a field of computer science that incorporates statistical techniques to provide a system the ability to automatically learn from past experience and improve how they execute certain tasks.

At the most fundamental level, classical machine learning can be categorized into three basic types: supervised learning, unsupervised learning, and semi-supervised learning. Additionally, there is another type of a machine learning technique called reinforcement learning (RL). Depending on the goal and the data a user has at its disposal, the type of algorithm changes. For instance, ***supervised learning*** is defined by its use of labeled data to train algorithms, which are later used to predict future outcomes. In contrast, ***unsupervised learning*** utilizes machine learning models to analyze and cluster a given dataset. The technique of ***semi-supervised learning*** falls in between the prior mentioned models since they use both labeled and unlabeled data for the training process. Last but not least, ***reinforcement learning*** is a behavioral model that does not rely on historical data but learns as it goes by using a trial and error approach (stimuli from an agent's environment).

3.2 Characteristics of Binary Classification Problems

The set-up for any **supervised learning** problem is as follows: by using labeled training data, learn a model that has the ability to predict the class of unseen and unlabeled test data. This illustrates also the clear contrast to **unsupervised learning** or clustering, where the target variable Y_i is unknown and the main goal is to detect hidden patterns in the data.

As Roth (2016) illustrates, consider a system that applies a function f to inputs x that returns an output $y = f(x)$. Given an **instance** or **feature space** X and a **label space** Y_i, which can be combined represented as $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where $X_i \in \mathbb{R}^p$ and $y_i \in Y_i$, there exists some target function $y = f(x)$, so that for any $x \in X$, this function outputs the correct y_i in the label space. Therefore, a supervised learner deals with a system where $f(x)$ is learned from a given set of (x, y) pairs, with the ultimate goal of finding a function $g(x)$ that is as close as possible to $f(x)$. While the random target variable, in case of a binary classification task, only take on two values $Y_i \in \{0, 1\}$ (where 0 is defined as non-default, and 1 corresponds to default), the feature space X can have both numeric and categorical inputs (see section 4 for more details).

The chosen classification methods in this work project are divided into two sections: (1) **Traditional scoring techniques** and (2) **advanced modeling methods**. For the first category, Logistic Regression as the most commonly used scoring model has been picked; for the second class, Decision Tree, Random Forest, XGBoost, AdaBoost, and Support Vector Machine were selected. The theory for these classifiers are explained in the succeeding subsections.

3.3 Traditional Modeling Technique: Logistic Regression

To explain the idea behind logistic regression as a model for binary classification, a quick refresher in the domain of probability theory must be made. Learning a binary classifier means to train a model of $p(y = 1|x)$ and $p(y = 0|x)$. Given that $p(y = 0|x) + p(y = 1|x) = 1$, it is sufficient to learn a model for only one of the class probabilities, for example $p(y = 1|x)$, from which $p(y = 0|x)$ will follow (Lindholm et al., 2021).

From the feature space definition of the previous section it is possible to re-write X_i in vector form as $x = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$ where p is the total amount of features in the dataset.

Simplifying this even further, one can write

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \boldsymbol{\beta}^T x. \tag{3.3.1}$$

Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic or sigmoid function to squeeze the output of a linear equation between the interval [0,1] (see Figure A.2). The logistic function $\phi(z)$ is defined as (Molnar, 2021):

$$\phi(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}} \tag{3.3.2}$$

As visible from Figure A.2, the sigmoid function is typically S-shaped and asymptotically bounded between 0 and 1. Furthermore, $\phi(z)$ approaches 1 if z reaches infinity ($z \to \infty$) since e^{-z} becomes small for large values of z. Similarly, $\phi(z)$ approaches 0 for $z \to -\infty$ as a result of an increasingly large denominator (Raschka and Mirjalili, 2019). Given this setting, this will make it possible to obtain a prediction of the respective class from the obtained z value. The more close inputs are to zero, and thus close to the decision boundary of $\phi(z) = 0.50$, the less capability the model has of making predictions of either class 0 or 1 for the given observation.

Based on the definition established in the introduction, the output of the sigmoid-function can then be interpreted as the probability of an instance belonging to class 1 $\phi(z) = p(y = 1|x)$, given its inputs x. For example, if the goal is to detect a potential loan default for a new customer, and the result is $\phi(z) = 0.70$, this would mean that the possibility of a loan default is 70%. As a result, the probability that the same client is not in danger of defaulting on his loan is $p(y = 0|x) = 1 - p(y = 1|x) = 0.30$ or 30%.

The predicted probability $\phi(z)$ can then be converted into a binary outcome via a threshold function (Raschka and Mirjalili, 2019):

$$\hat{y} = \begin{cases} 1, & \text{if } \phi(z) \geq 0.5 \\ 0, & \text{otherwise.} \end{cases} \tag{3.3.3}$$

Putting Equation 3.3.1 and 3.3.2 together, it is possible to obtain a generalized function for the interval [0,1], which can be used to model the outcome $p(y = 1|x)$,

$$p(y = 1|x) = \frac{e^{\boldsymbol{\beta}^T x}}{1 + e^{\boldsymbol{\beta}^T x}} \tag{3.3.4}$$

Rearranging Equation 3.3.4 yields the logit transformation (which is where logistic regression gets its name):

$$g(x) = \log\left[\frac{\phi(x)}{1-\phi(x)}\right] = \boldsymbol{\beta}^T x \tag{3.3.5}$$

Applying this logit transformation to $\phi(x)$ results in a linear equation similar to the OLS linear regression model. The term in the $log()$ is called *odds* and in combination with the log it is also referred to as *log odds*.

The only unknown parameters in this equation are $\boldsymbol{\beta}$ which will be learned from the training data. The estimation is performed by using the cost function ***maximum likelihood method*** (MLE). The estimators $\boldsymbol{\beta}$ are chosen to maximize Equation 3.3.6.

$$\ell(\boldsymbol{\beta}) = \prod_{i:y_i=1} p(X_i) \prod_{i':y'_i=0} \left[1 - p\left(x'_i\right)\right] \tag{3.3.6}$$

Maximum likelihood tries to find $\boldsymbol{\beta}$ in such a way that the predicted probabilities are as close to the observed probabilities as possible. For a binary classification, MLE will try to find values of the $\boldsymbol{\beta}$ such that the resultant probabilities are closest to either 1 or 0 (Saraswat, 2017).

3.4 Advanced Modeling Techniques

The biggest problem with traditional credit scoring based on logistic regression is that as a practitioner or researcher, one cannot interpret the importance of underlying variables to the probability of a borrower experiencing financial difficulty (Sharma, 2011). Despite the fact that p-values can signal the relative importance of individual features, they will however become unreliable when multicollinearity (or collinearity) comes into play since high correlations among predictor variables may lead to unreliable and unstable estimates of regression coefficients. This constitutes a severe problem, particularly for credit risk data since certain socioeconomic variables, such as income, age and education are highly correlated.

To overcome those obstacles, advanced statistical models may provide a more scientific approach to analyze feature importance and to achieve higher predictive accuracy. As a result, the following sub-chapter briefly presents a overview of the utilized (supervised) classification models that are used as comparison classifiers against logistic regression. However, given the mathematical complexity of these algorithms, only a textual summary of the core fundamental concepts are highlighted.

3.4.1 Tree models

Decision Tree and Random Forest, as representatives of tree-based models, are a class of non-parametric algorithms that work by partitioning the feature space into a number of smaller (non-overlapping) regions with similar response values using a set of splitting rules (Boehmke and Greenwell, 2020). The partitioning or splitting is based on learning simple decision rules inferred from the training data. Moreover, from a computational point of view, trees are regarded as being not only highly efficient given their performance on larger datasets but also they require less data preparation than other classification techniques.

3.4.2 Boosting models

Boosting is part of the ensemble family which constitutes another approach to improve the predictions resulting from a decision tree. As Lindholm et al. (2021) emphasize, boosting is built on the idea that even a simple (or weak) high-bias classification model often can capture some of the relationships between the inputs and the output. Boosting works in a similar way as bagging, except that the trees are grown sequentially, while in bagging the results of multiple classifiers are combined and modeled on different sub-samples of the same dataset. Unlike Random Forest, boosting does not involve bootstrap sampling; instead, each tree is fitted on a modified version of the original dataset (Le, 2018). Similar to bagging, boosting can be applied to any type of model, however, most often it is used in combination with Decision Trees. A general framework for boosting can be seen in Figure A.8.

3.4.3 Support Vector Machine

The Support Vector Machine (SVM) method is a set of supervised learning technique, able to solve both regression and classification tasks. The model, which incorporates statistical learning theory, was first introduced by Vapnik (1998) and it constitutes a global, constraint, optimized learning algorithm based on Lagrange multipliers method (Khemakhem and Boujelbene, 2017). Today, there are several successful application areas of SVM, ranging from image recognition to text classification and computational biology.

For a binary classification problem, SVM seeks to separate the two classes by mapping the input vectors to a higher dimensional space and then constructing an optimal hyperplane with maximum margin to achieve optimal separation (margin) between the classes (Kaya et al., 2008). For better understanding, the theoretical concept of SVM is highlighted in Figure A.9.

3.5 Model Evaluation Techniques

3.5.1 Confusion matrix

For a binary classification problem, most often, a confusion matrix is used to evaluate the performance of a model. It determines how many observations were correctly or incorrectly classified - thus, it provides a more detailed breakdown of the results for each class. In a binary context, it is common to present the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions (see Figure A.3 for an exemplary illustration).

3.5.2 Binary Classification Measures

The most straightforward classification metric in machine learning is considered to be ***accuracy***. It is defined as the percentage of the correctly classified observations, as shown by the formula:

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \tag{3.5.1}$$

An alternative metric to accuracy is ***precision***. It is the fraction of instances predicted as positive divided by the actual positives. In other words, precision answers the question of what proportion of positive predictions were actually correct. The mathematical notation is:

$$Precision = \frac{TP}{FP + TP} \tag{3.5.2}$$

Recall (or sensitivity) measures how many positive instances are correctly predicted amongst all positive samples. Formally, it is the ratio between the true positive rate (TPR) and the sum of the correctly predicted labels (TP) and wrongly predicted classes (FN).

$$Recall = \frac{TP}{FN + TP} \tag{3.5.3}$$

Furthermore, during the hyperparameter-optimization process, the recall score is used as an input to leverage the best results. When the cost of having false negative is high, the aim of a financial institution should be to increase a model's sensitivity since the downside-risk would be to incorrectly label a solvent customer as a potential loan-defaulter.

3.5.3 Receiver Operating Characteristic (ROC)

A ROC-curve, short for *receiver operating characteristic curve*, is a common graphical plot to visualize the performance of a binary classifier at all classification thresholds. While the x-axis of the ROC-curve shows the false positive rate, the y-axis displays the true positive rate. Whereas classifiers that are closer to the top-left corner indicate a better relative performance, models that are closer to the 45-degree diagonal of the ROC space indicate less accurate algorithmic performance (for example, compare the relative performance of classifiers A and B in Figure 3.5.1). The optimization is performed by maximizing the true positive rate while minimizing the false positive rate (Pedregosa et al., 2011). Additionally, as a baseline, a random classifier is expected to give points lying along the diagonal, where FPR = TPR (Chan, 2020).

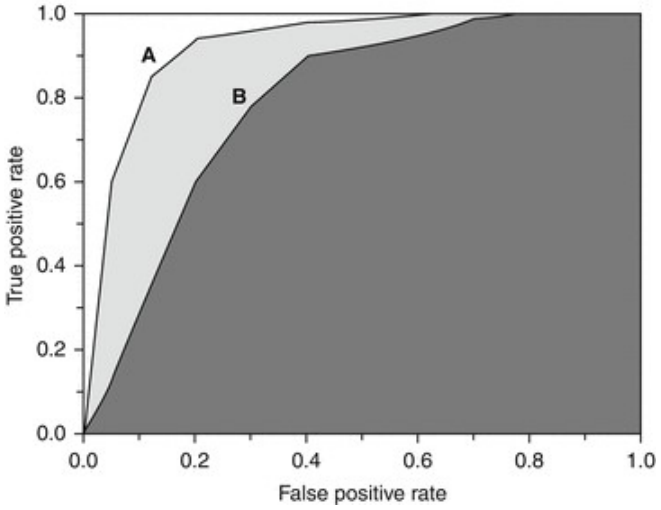

Figure 3.5.1: ROC-curve [Melo (2013)]

3.5.4 Area Under the Curve (AUC)

The *area under a receiver operating characteristic curve*, abbreviated as AUC, is a single aggregated scalar that measures the classification performance across all possible thresholds of an algorithm by calculating the area underneath the ROC-curve, similar to an integral ranging from the origin (0,0) to (1,1). One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example (Moura Oliveira et al., 2019). As seen from Figure 3.5.1, classifier A has a larger AUC-value in comparison to classifier B, given that its ROC-curve is closer to the top left corner of the graph.

Chapter 4

Data Management and CR-Pipeline

In this section, a description of the dataset, basic pre-processing steps, a variable selection technique, and the introduction of the credit risk-pipeline will be discussed.

4.1 Dataset characteristics

In this work project, a publicly available open-source credit risk data from Kaggle, published by Lao Tse, was used to model customer loan defaults. The data contains basic financial information of 32,581 borrowers with a total of 12 attributes regarding a multitude of client-specific characteristics and loan contract details. A complete overview of the explanatory variables, in combination with a short description, can be seen in Table B3. In addition, descriptive statistics are displayed in Table B4.

The binary outcome variable *status* indicates if a customer defaulted on its obligation or not (0: non-default, 1: default). Conducting a summary statistic on the target feature yields the result of an unbalanced or highly skewed dataset with 78.18% non-defaults and 21.82% defaults. Given the nature of the problem, this is not surprising since repaying a loan is more likely than defaulting. However, class imbalance is a concerning problem with regard to machine learning because the error rate is insensitive to the classification accuracy of models. Therefore, the proposed algorithms will struggle to learn during the training process, given that they are not provided with enough relevant data. To overcome this issue, a methodology called Synthetic Minority Oversampling (SMOTE) that helps to tackle this problem, is introduced in section 4.4.

4.2 Data preprocessing

By far the most time-consuming and important work next to the actual modeling in machine learning is data preparation. According to a survey conducted by *CrowdFlower* in 2016, data pre-processing accounts for approx. 80% of the work of data scientists (Press, 2016).

Essentially, data preparation refers to the task of preparing or transforming raw data to make it suitable and understandable for statistical models. As Zhou et al. (2017) highlight, the following common pre-processing tasks can be distinguished: removal of duplicates & missing values, outlier detection, discretization of data, feature selection, and data labeling. Each of those mentioned operations aims to help the model to achieve better predictive power. Since the details of the individual pre-processing procedures are outside the scope of this paper, the main focus will be laid on a more in-depth analysis of the constructed data preparation pipeline, which combines multiple processing steps together (see section 4.2.3 for more details).

4.2.1 Outlier & missing value handling

As seen from Table B4, the dataset contains outliers as well as missing values. For example, anomalies can be detected in the variables *age*, *employment length* and *income*. Conducting an exploratory data analysis (EDA), this can also be demonstrated by a distribution plot (shown for *age* in Figure A.4 and for *employment length* in Figure A.5).

Regarding missing values in the data, for only two attributes, *interest rate* and *employment length*, null-values were found. Given their insignificance based on the amount (9.6% respectively 2.7%), the missing values were handled via median-imputation as part of the credit risk pipeline (see section 4.2.3 for more details). After outlier-removal, the dataset was reduced from 32,581 to 31,679 rows, constituting a deletion of a total of 902 records. However, depending on the nature of missing values (MCAR, MAR, and MNAR processes), even a highly sophisticated credit risk model might fail if the data used for modeling purposes is inadequate. In the common literature, this phenomenon is described as *selection bias* since the subset used for modeling is not a representative choice of the overall dataset (Little and Rubin, 1991). Hence, the characteristics of the incomplete data must be checked to ensure that no abnormal behaviour is incorporated in this sample. By comparing the original data (Table B5) with the missing value dataset (Table B6), no pattern in the null values can be detected that would deduce unusual behaviour on a broad scale.

4.2.2 Correlation analysis

To summarize the dependencies between the variables, a correlation matrix as a diagnostics tool for more advanced analyses was computed (for graphical enhancement, a heatmap is used as a substitute), where high positive correlation is shown in dark-blue, high negative correlations in bright-yellow, while light-green represents no correlation between two features.

As seen from Figure A.6, strong positive correlation can be detected between the variables *age* and *credit history length* (0.88), and between *amount* and *loan to income ratio* (0.57). In both cases, the dependency is expected and justifiably: the higher the age of a customer, the more likely it is that this person has an account with the bank for a longer time period. Also, the larger the loan amount, the higher the ratio in comparison to the income of a client. As highlighted, correlated features might constitute an issue when it comes to machine learning modeling. To overcome this problem, feature selection techniques will be discussed in section 4.3, which specifically can deal with dependencies among attributes.

4.2.3 Pipeline

Data cleaning and preparation constitutes the most time-consuming part of any machine learning project. Thus, to simplify this process, **pipelines** are a simple way to keep the data preprocessing and modeling code as organized and structured as possible. In essence, a pipeline bundles preprocessing and modeling steps together as if it would be one single step. Most often, pipelines include the execution of the following tasks: data normalization (scaling), imputation of missing values, dimensionality reduction, dummy, and categorical encoding. A generic overview of the pipeline utilized in this project can be seen in Figure 4.2.1.

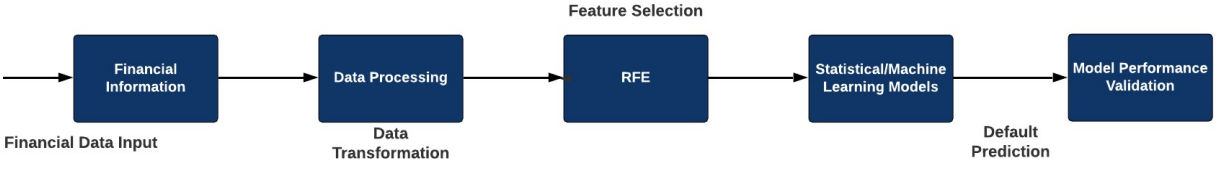

Figure 4.2.1: Financial credit risk assessment pipeline

The individual components that were used within the pipeline are part of the *scikit-learn.preprocessing* package. *Scikit-learn* is an open-source machine learning library, and undoubtedly the gold standard for machine learning tasks in the Python ecosystem. It supports not only supervised but also unsupervised learning and various other tools for model pre-processing, variable selection, and evaluation.

4.3 Feature selection

Variable selection is one of the core concepts in machine learning, where the goal is to reduce the number of attributes from a given set of features according to certain criteria. By discarding certain variables, noticeable effects may emerge: the performance of prediction models might improve, the training time will be reduced and the remaining data will be less prone to overfitting, given that there are fewer opportunities to make decisions based on noisy data (Liu et al., 2010). In this project, a (wrapper) selection technique called *Recursive Feature Elimination* (RFE) is utilized. As Guyon et al. (2002) explain, RFE is regarded as a backward selection technique for the predictors. The basic idea behind this method is that it starts by building a model on the entire dataset, and simultaneously assign weights or feature importance scores to each of the attributes according to a pre-defined external estimator. Based on those scores, the least important predictor(s) are removed recursively, the model is re-built, and feature importance's scores are re-computed again (Kuhn and Johnson, 2019). That procedure is recursively repeated on the down-scaled dataset until a desired number of attributes is reached (Pedregosa et al., 2011).

4.4 Treatment of imbalanced data with SMOTE

By definition, imbalanced classification constitutes a common machine learning problem where the majority class outnumbers the minority class by a large proportion (Ling and Sheng, 2011). This particularly becomes an issue for predictive models that assume equal class distribution by default (Brownlee, 2021). If not handled properly, the performance of such classifiers on the minority class will deteriorate, given the bias towards the majority class (Akosa, 2017).

A popular extension to the *oversampling* algorithm, proposed by Chawla et al. (2002), constitutes **S***ynthetic* **M***inority* **O***versampling* **T***echnique* (SMOTE), which generates synthetic samples from the minority class. The general idea behind this method is that instead of deleting examples from the majority class or duplicating observations from the underrepresented class like *undersampling* respectively *oversampling*, SMOTE overcomes the data imbalance by generating artificial data based on linear interpolation between minority class samples and their K nearest neighbors (Lv et al., 2019). The method, as described by Hu and Li (2013), works as follows: First, for each minority class sample x, one gets its k-nearest neighbors from other minority class samples. Second, one chooses one minority class sample $\bar{x}$ among the k neighbors. In the last step, this generates the artificial sample $\mathrm{x}_{\mathrm{new}}$ by interpolating between x and $\bar{x}$, as pictured in Figure A.11.

Chapter 5

Results

In this chapter, the obtained results from the previously mentioned models and their feature importance will be presented and discussed. Furthermore, a practical use-case example of loan default prediction via machine learning will be given.

5.1 Performance of the overall models

As mentioned, variable selection was performed via RFECV; in total 16 features were selected (see Figure A.10), where minority oversampling with SMOTE was applied. In addition, the overall performance of the models is evaluated on multiple criteria, foremost on accuracy, recall/sensitivity, and precision. However, special attention is also given to the estimates of the confusion matrix, particularly to TNs and FNs. The combined results are presented in Table 5.1. Additionally, to show the visual performance of the algorithms, ROC curves were plotted (see Figure A.12).

Model	**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**TPs (%)**	**FPs (%)**	**FNs (%)**	**TNs (%)**
Logistic Regression	80.54	78.55	53.27	16.92	14.84	4.62	63.62
XGBoost	92.42	77.10	86.27	16.61	2.64	4.93	75.82
AdaBoost	85.15	76.97	62.63	16.58	9.89	4.96	68.57
Decision Tree	76.63	82.83	47.56	17.48	19.67	3.70	58.79
Random Forest	91.07	75.93	81.37	16.35	3.74	5.18	74.72
Support Vector Machine	21.54	100.00	21.54	21.54	78.46	0.00	0.00

Table 5.1: Model performance summary on train/test-set

From Table 5.1 it can be noticed that the outcomes of the RF and XGB-classifier are very much alike in terms of their overall positive performance, while the SVM was not able to capture any false negatives nor true negatives of the provided dataset. Interestingly, while LR as the

traditional credit risk classifier did not perform the best out of the mentioned models, it did however provide similarly (sometimes even better) results than certain machine learning algorithms. Furthermore, a general conclusion can be drawn that the boosting methods (XGBoost and AdaBoost) showed on average better results than the tree-based models (Decision Tree and Random Forest).

To receive a visual impression of the relative performance of the individual models, ROC curves were created, supporting the interpretation and analysis from above. From Figure A.12, similar to the numeric evaluation, it can be concluded that there is no worth mentioning difference in the ROC curves and AUC scores between the best performing models (LR, XGB, and RF). Note that the black dashed diagonal line in the ROC plot represents a dummy classifier with 50% sensitivity and 50% specificity (Zhu et al., 2010).

5.2 Results of the best performing models on the holdout set

In the previous section, the performance of the models was measured solely on the training and test data. However, the goal of a well-trained artificial intelligence algorithm is not only to perform well during the training process but also to generalize on an unseen validation set. If a model works well on the training data but fails to generalize on the validation/holdout set, it is said that the model is **overfitting** or sometimes also referred to as memorizing, given that instead of learning the distribution of the data, it simply learns the expected output for every data point. The opposite is called **underfitting**, which is the event where an algorithm can neither model the training data nor generalize to unseen data (Brownlee, 2021).

Therefore, to measure the generalization of the best performing models from the last section (XGB and RF) on a holdout set, their performance metrics are displayed in Table 5.2.

Model	**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**TPs (%)**	**FPs (%)**	**FNs (%)**	**TNs (%)**
XGBoost	92.05	74.61	86.62	16.08	2.48	5.47	75.97
Random Forest	90.36	72.27	80.96	15.57	3.66	5.98	74.79

Table 5.2: Model performance summary on validation set

Since the key metrics of the holdout set are in line with the estimates of the training/test process, it is possible to conclude that the classifiers neither over- nor underfit on unseen data and that the XGB as well as the RF algorithm yield stable outcomes for new predictions.

5.3 Model interpretation with SHAP

Most often, a well-trained and sophisticated machine learning algorithm may produce reliable and accurate predictions, but notoriously lacks when it comes to feature interpretation since many methods, for example, XGB, are regarded as so-called *black-box models*. However, when building a model - like in this case, a procedure that can predict potential loan defaulters, it is essential for the end-user to understand not only which variables have an impact on the final decision, but also how it is making the prediction. To bridge this gap between models that are accurate but difficult to interpret, a novel approach called SHAP (SHapley Additive exPlanations) is proposed.

Armed with this new methodology, specifically with the **global importance plot** that is incorporated within the Shapley value framework, interpreting the impact of individual features becomes a straight-forward task. The idea behind this method is simple: the global importance plot sums up the absolute Shapley values per feature across the data, where attributes with larger absolute values are considered as more important.

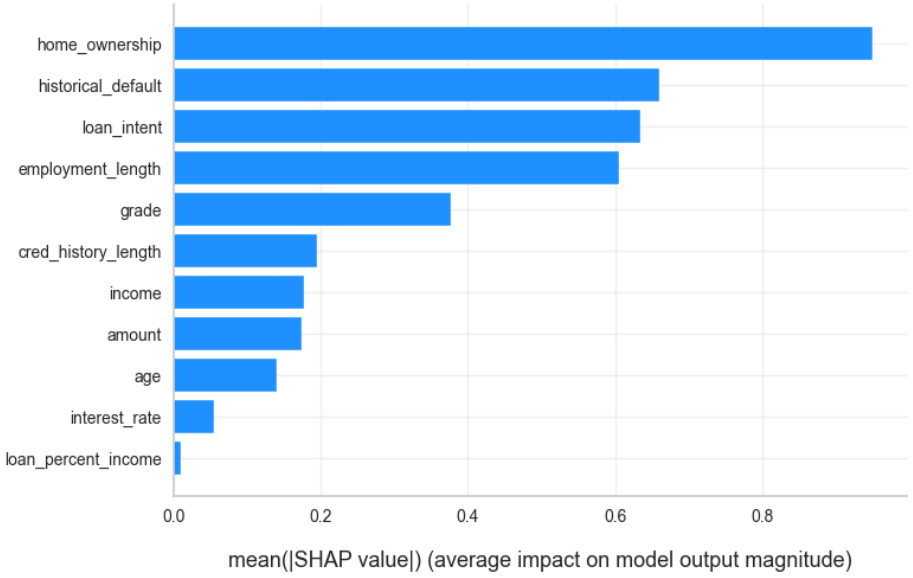

Figure 5.3.1: SHAP (global) feature importance plot

The variable importance plot in Figure 5.3.1 lists the most significant features in descending order. Thus, the former black-box XGB classifier can be interpreted as follows: to predict future loan defaults, a moneylender should consider the home-ownership status, the customer's history of past loan defaults and the client's loan intent as the most significant attributes. For instance, given that *home_ownership* is the most important feature, it would change the predicted absolute default probability on average by 94.92 percentage points (0.949217 on the x-axis).

5.4 Application of the credit risk model

In the following, a simplified version of a graphical user interface (GUI) that can predict potential future loan defaults is shown. The GUI was written with the popular ipywidgets library that interactively displays HTML widgets for Jupyter notebooks and the IPython kernel.

Based on the tool that is shown in Figure 5.4.1, a money-lending institution can plug in multiple inputs like age, income, home-ownership status, employment-length, loan intent, credit-grade, interest rate, loan-to-income-ratio, historical-default status, credit history length and loan-amount, to assess the credit risk of a customer. Furthermore, the user is able to choose between two classifiers, which are *XGBoost* and *Random Forest*.

Practical use-case example:

Plugging in the following inputs for a hypothetical client called Francisco, who is 30 years old, has an average yearly income of 63,000 USD, lives in a rented dwelling, works since at 8 years, needs the loan for home-improvement purposes, has a credit rating of grade B, a loan-to-income ratio of 15%, did not experience a loan default in the past, a credit history length of 5.5 years, and who asks for a total amount of 11,000 USD, will have a low probability of default (PD of approx. 7.78%), according to the XGBoost algorithm, as seen from Figure 5.4.1.

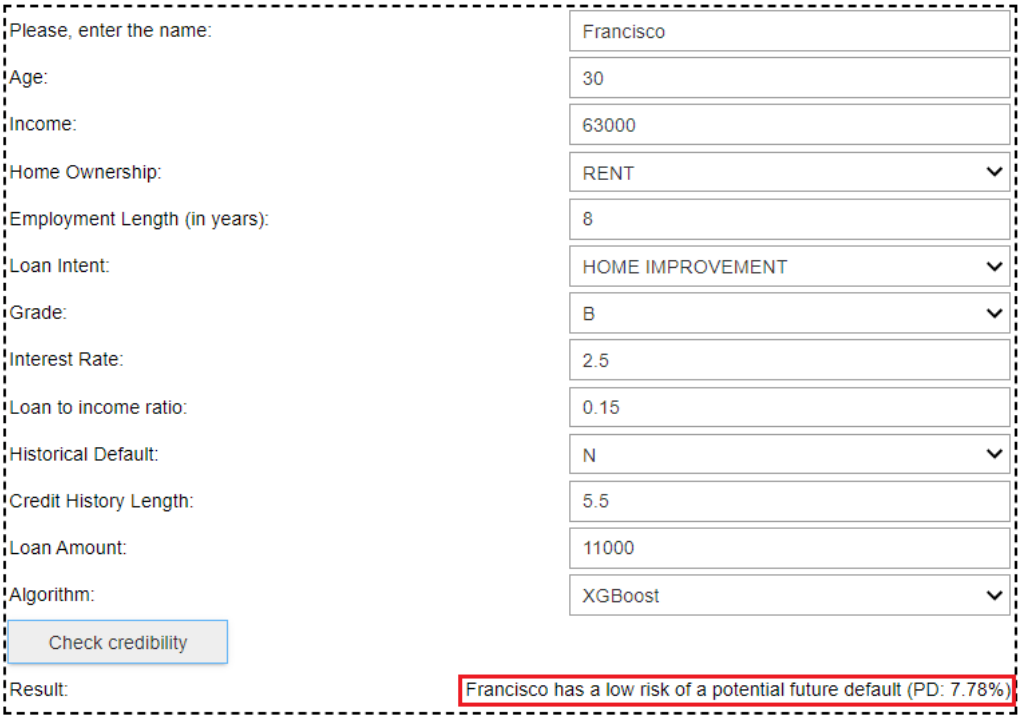

Figure 5.4.1: Credit risk prediction by the GUI

Chapter 6

Discussion

In the final section of this thesis, the outcomes presented in the previous section will be discussed and analyzed, the posed research questions answered and possible limitations briefly addressed.

6.1 Research questions

In this thesis, a comparison was conducted between traditional credit risk models used in the industry, and more advanced, state-of-the-art machine learning algorithms. The first goal was to find an answer to the question if machine learning-based credit assessment models are better able to predict customer loan defaults than traditional empirical models. Finally, the second goal dealt with the question, which attributes have the largest impact on the final model.

The answers to the research questions are as follows:

1. ***If machine learning based classifiers are better able to predict applicants' loan defaults than traditional empirical models, how well do they perform in their assessment?***

In section 5 of this paper, a comparison was conducted between LR as the representative of the traditional models, and multiple machine learning methods as the proxy for the more advanced statistical techniques to determine, if a significant difference in the classification performance can be detected. It was possible to observe that well-trained machine learning models like XGB and RF were able to outperform LR in several key metrics on a constant basis (for instance, accuracy, recall, FN, and FP-rate). Hence, one can conclude that machine learning-based credit

risk methods are better in predicting customer loan defaults than traditional credit assessment models.

2. ***Which credit variables/features are the key drivers for customer loan defaults?***

This question was addressed in section 5.3 of this thesis. The approach used to obtain the most significant attributes is the so-called SHAP value method. Based on this game-theoretical methodology, the top three features that inherit the largest predictive explainability are the home-ownership status, the customer's history of past loan defaults, and the client's loan intent.

6.2 Limitations and future work

The future tasks with respect to credit risk models via machine learning algorithms are:

- The progress on computers' processing capabilities and the widespread availability of suitable data allows the use of methods that are regarded as even more advanced than supervised machine learning techniques, such as Deep Learning models, with the most prominently used algorithm within this framework - *Artificial Neural Nets*. Given its structure, which resembles a biological neural network, it may unveil new determinants and unexpected forms of dependencies among the features.

- The ML models were only trained and tested against the standard default-probability threshold equal to 0.5 (probabilities above 0.5 mean default, values below non-default). Future work may involve threshold-tuning to obtain more conclusive and realistic results.

The future tasks with respect to model explainability:

- As highlighted in section 5.3, SHAP was primarily used for interpreting the best performing model. However, multiple other libraries within the Python framework exist, for example, *LIME* and ELI_5, each one of them incorporating a different unique approach to explain and interpret the decisions of a predictive model.

- The conducted feature importance analysis in chapter 5.3 was merely tested using the XGBoost model. Alternatively, other algorithms with similar classification results, for instance, Random Forest, could be used for the explainability and interpretability of the model's outcome to enable further transparency in the decision-making process.

Appendix

A Figures

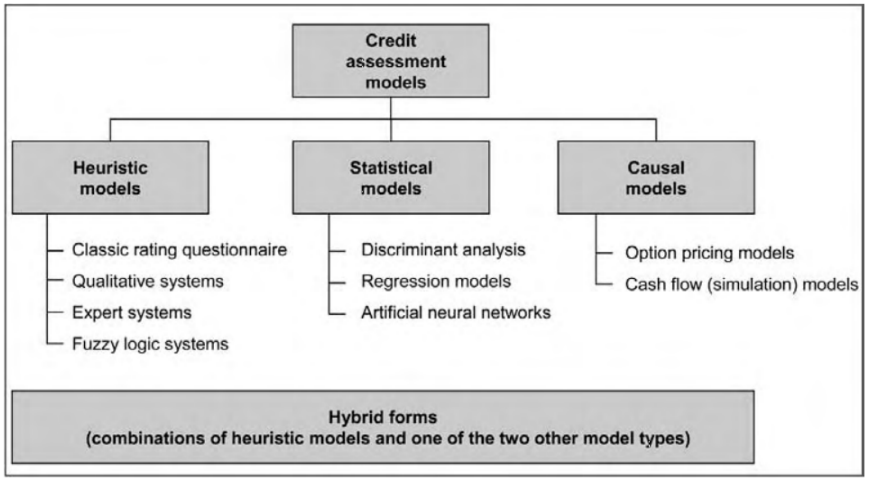

Figure A.1: Overview of Credit Assessment Models [OeNB (2004)]

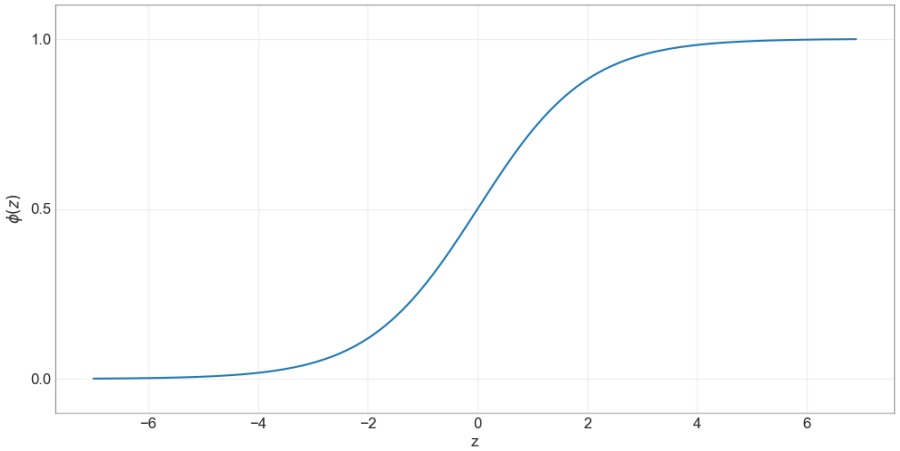

Figure A.2: Sigmoid Function $\phi(z)$

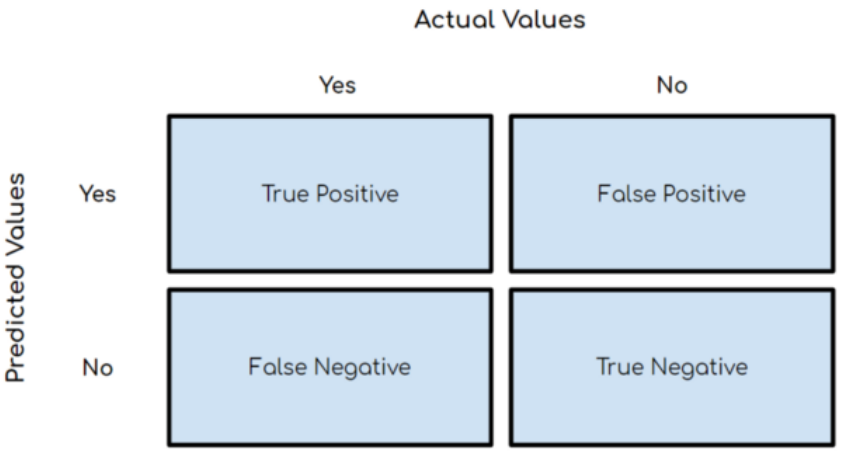

Figure A.3: Binary Confusion Matrix [Shin (2020)]

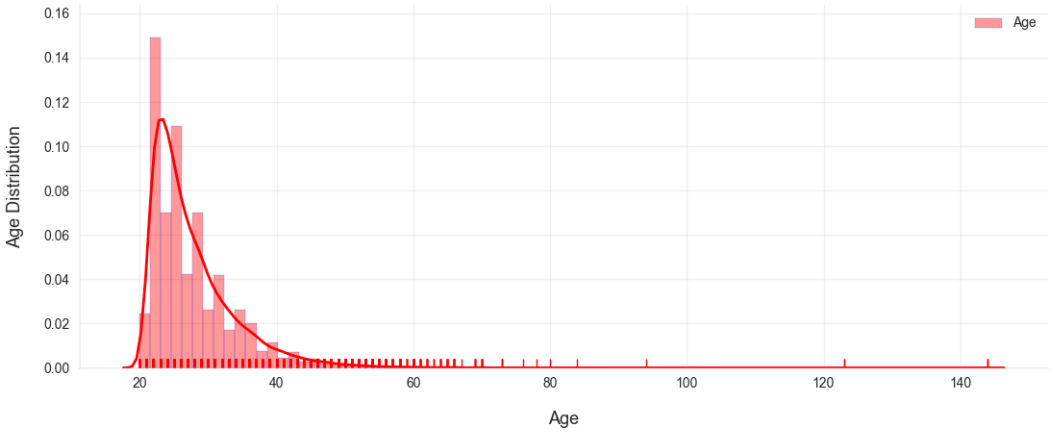

Figure A.4: Distribution of loan applicant's age

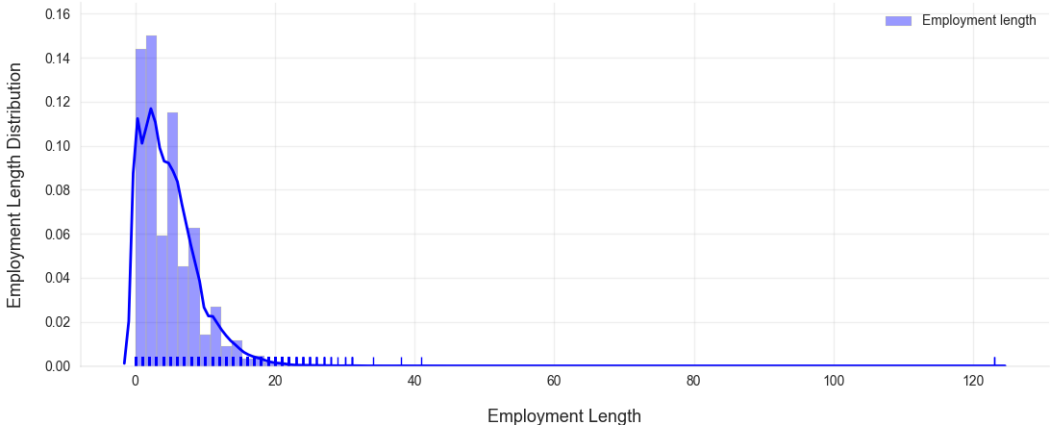

Figure A.5: Distribution of loan applicant's employment length

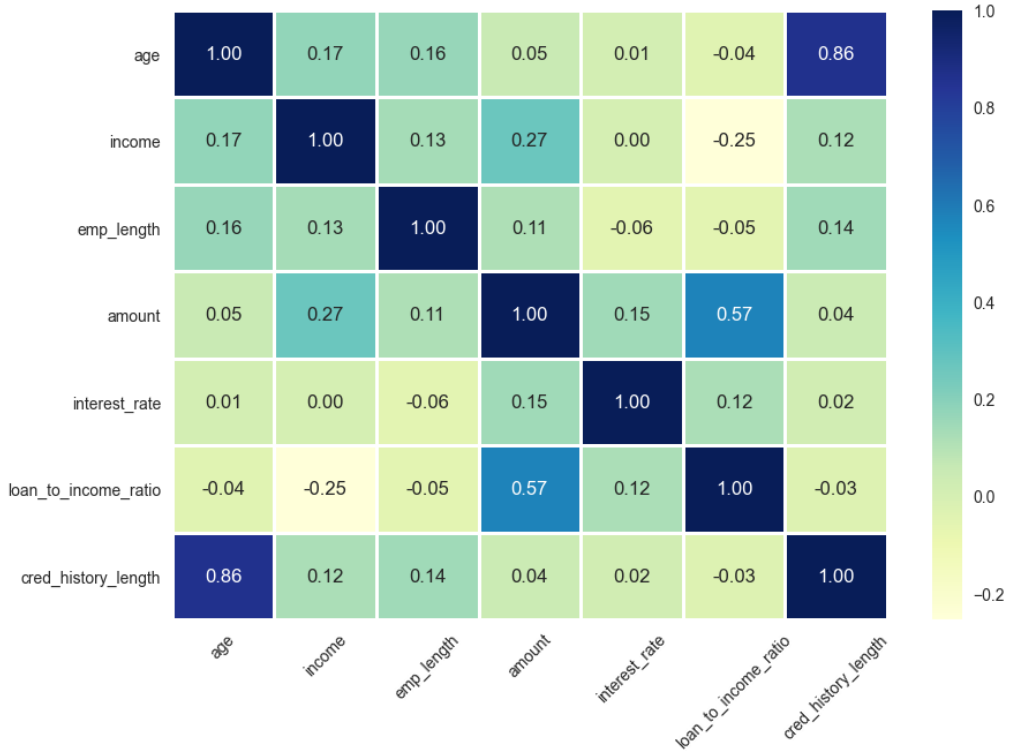

Figure A.6: Correlation matrix (heatmap)

```
    numeric_transformer = Pipeline(steps=[
        ('imputer', SimpleImputer(strategy='median')),
        ('scaler', StandardScaler())])

    categorical_transformer = Pipeline(steps=[
        ('imputer', SimpleImputer(strategy='constant', fill_value='missing')),
        ('onehot', OneHotEncoder(handle_unknown='ignore', sparse=False))])

    numeric_features = X.select_dtypes(include=np.number).columns
    categorical_features = X.select_dtypes(exclude=['number']).columns

    preprocessor = ColumnTransformer(
        transformers=[
            ('num', numeric_transformer, numeric_features),
            ('cat', categorical_transformer, categorical_features)])

    preprocessor.fit(X_train)

    # Get final feature names
    cat_columns = preprocessor.named_transformers_['cat']['onehot'].get_feature_
    names(categorical_features)
    columns_pipeline = np.append(cat_columns, numeric_features)

```

Figure A.7: Pre-processing pipeline

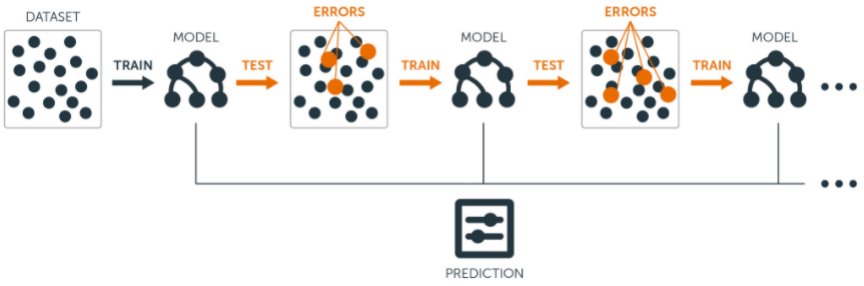

Figure A.8: Boosting framework [Le (2018)]

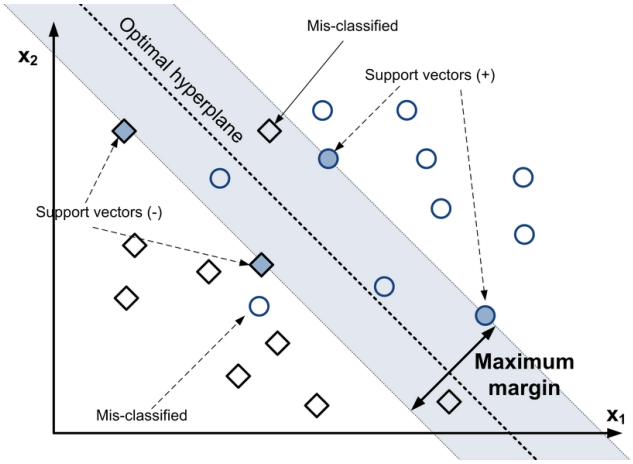

Figure A.9: Concept of Support Vector Machines [Nguyen Duc et al. (2017)]

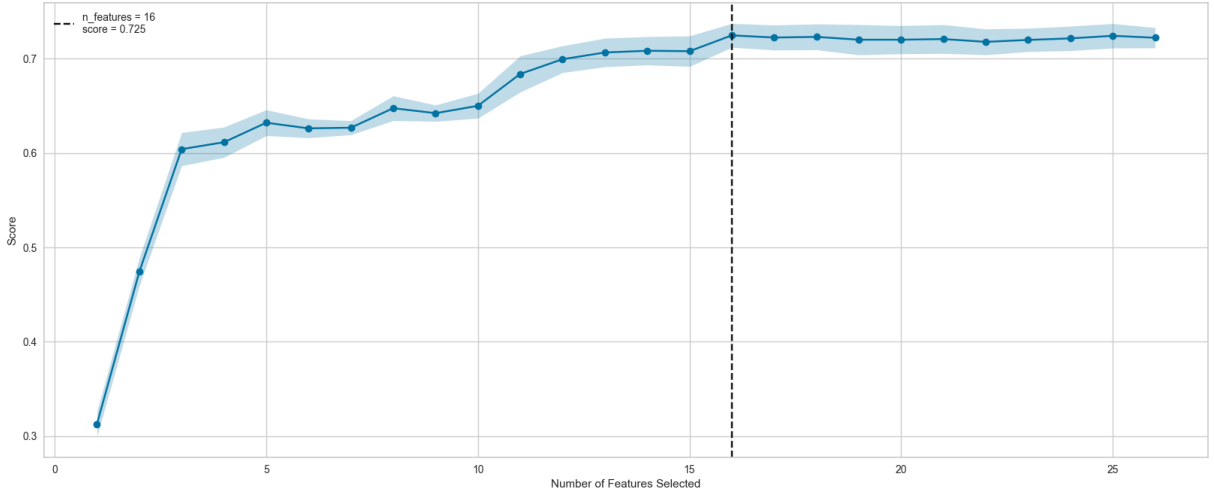

Figure A.10: RFECV for RandomForestClassifier

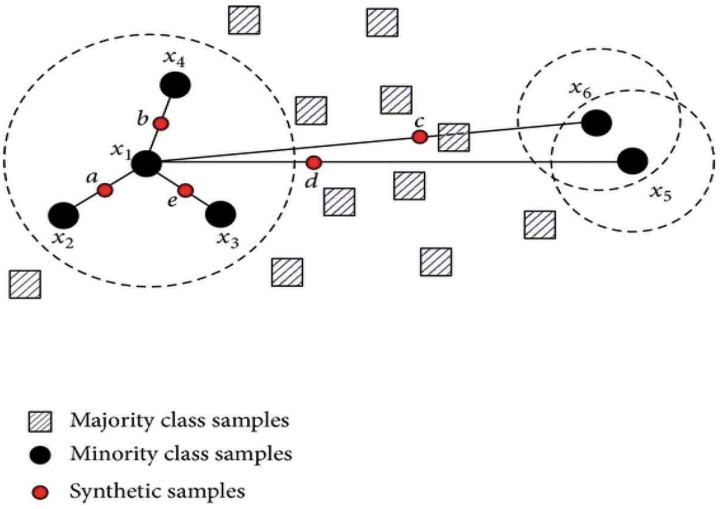

Figure A.11: SMOTE algorithm [Shrivastava et al. (2020)]

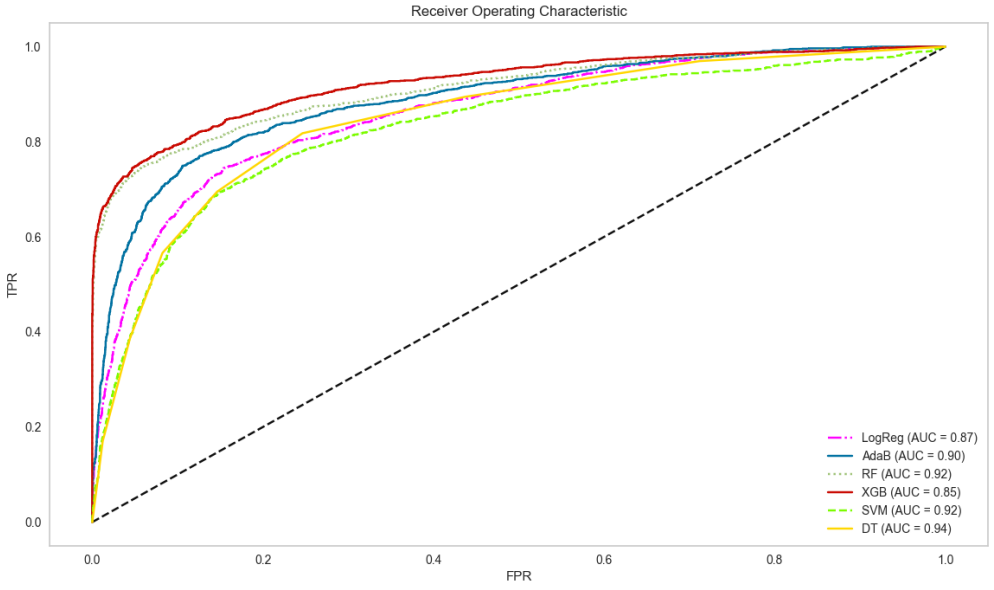

Figure A.12: ROC curves for RFECV selected feature models

B Tables

Risk Parameter	Foundation IRB	Advanced IRB
Probability of default	Estimated by the bank	Estimated by the bank
Loss Given Default	Supervisory value	Estimated by the bank
Exposure at Default	Supervisory value	Estimated by the bank

Table B1: Overview of risk parameters per model

Year	Technique	Work
1996	Statistical methods	Dimitras et al. (1996)
1997	Statistical methods	Hand and Henley (1997)
1998	Neural networks	Wong and Selvi (1998)
1999	Neural networks	Vellido et al. (1999)
2000	Statistical and operational methods	Thomas (2000)
2002	Neural networks	Calderon and Cheh (2002)
2006	Statistical and machine learning methods	Balcaen and Ooghe (2006)
2007	Statistical and machine learning methods	Kumar and Ravi (2007)
2010	Hybrid and ensemble techniques	Verikas et al. (2009)
2010	Machine learning methods	Paleologo et al. (2010)
2010	Machine learning methods	Twala (2010)
2011	Support vector machines	Jayanthi et al. (2011)
2012	Neural computing	Brabazon et al. (2012)
2012	Machine learning methods	Lin et al. (2012)
2013	Ensemble techniques	Woźniak et al. (2013)
2014	Semi-parametric methods	Lam and Trinkle (2014)
2016	Non-parametric methods	Deng (2016)
2017	Machine learning methods	Akindaini (2017)
2017	Machine learning methods	Hué et al. (2017)
2018	Machine/Deep learning methods	Addo et al. (2018)

Table B2: Results from selected papers

Variables	**Type**	**Detail**
age	Numeric	Is the age of the person at the time of the loan
income	Numeric	Is the yearly income of the person at the time of the loan
home status	Categorical	Is the type of ownership of the applicant (“rent”, “mortgage” or “own”)
employment length	Numeric	Is the amount of time in years the applicant is employed
loan intent	Categorical	Is the aim of the loan (“education”, “medical”, “venture”, “home improvement”, “personal” or “debt consolidation”
credit history length	Numeric	Tells how long ago a customer has opened up a bank account
historical default	Categorical	Answers whether the person has defaulted before (”Y” or ”N”)
loan to income ratio	Numeric	Is the ratio between the loan taken and the annual income
interest rate	Numeric	Is the interest paid for the loan expressed in percentage
loan grade	Categorical	Is a classification respectively loan scoring system used by financial institutions to grade loan applicants given their credit history, likelihood of repayment of interest and principal and quality of the collateral (”A”, ”B”, ”C”, ”D”, ”E”, ”F” and ”G”)
loan amount	Numeric	Is the loan taken dollar-amount
loan status	Categorical	Target variable (0: non-default, 1: default)

Table B3: Variable description of the dataset

variables	**count**	**unique**	**top**	**freq**	**mean**	**std**	**min**	**max**
age	32,581.00	-	-	-	27.73	6.35	20.00	144.00
income	32,581.00	-	-	-	66,074.85	61,983.12	4,000.00	6,000,000.00
home_ownership	32,581.00	4	RENT	16446	-	-	-	-
emp_length	31,686.00	-	-	-	4.79	4.14	0.00	123.00
loan_intent	32,581.00	6	EDUCATION	6453	-	-	-	-
grade	32,581.00	7	A	10777	-	-	-	-
amount	32,581.00	-	-	-	9,589.37	6,322.09	500.00	35,000.00
interest_rate	29,465.00	-	-	-	11.01	3.24	5.42	23.22
status	32,581.00	-	-	-	0.22	0.41	0.00	1.00
loan_percent_income	32,581.00	-	-	-	0.17	0.11	0.00	0.83
historical_default	32,581.00	2	N	26836	-	-	-	-
cred_history_length	32,581.00	-	-	-	5.80	4.06	2.00	30.00

Table B4: Descriptive summary statistic of the original dataset

	Default (mean)	Default (top)	Non-default (mean)	Non-default (top)
age	27.47	-	27.81	-
income	49,125.65	-	70,804.36	-
home_ownership	-	RENT	-	MORTGAGE
emp_length	4.14	-	4.97	-
loan_intent	-	MEDICAL	-	EDUCATION
grade	-	D	-	A
amount	10,850.50	-	9,237.46	-
interest_rate	13.06	-	10.44	-
status	1.0	-	0.0	-
loan_to_income_ratio	0.25	-	0.15	-
historical_default	-	N	-	N
cred_history_length	5.69	-	5.84	-

Table B5: Summary statistic of the original dataset

	Default (mean)	Default (top)	Non-default (mean)	Non-default (top)
age	27.64	-	27.86	-
income	42,750.43	-	67,747.55	-
home_ownership	-	RENT	-	MORTGAGE
emp_length	-	-	-	-
loan_intent	-	MEDICAL	-	EDUCATION
grade	-	B	-	A
amount	9,668.84	-	8,959.73	-
interest_rate	-	-	-	-
status	1.0	-	0.0	-
loan_to_income_ratio	0.25	-	0.15	-
historical_default	-	N	-	N
cred_history_length	5.76	-	5.93	-

Table B6: Summary statistic of the missing-value dataset

Bibliography

P. M. Addo, D. Guegan, and B. Hassani. Credit Risk Analysis Using Machine and Deep Learning Models. *Risks*, 6(38), 2018. doi: 10.3390/risks6020038.

B. Akindaini. Machine Learning Applications in Mortgage Default Prediction. Master's thesis, University of Tampere, November 2017. URL `https://trepo.tuni.fi/bitstream/handle/10024/102533/1513083673.pdf?sequence=1&isAllowed=y`.

J. S. Akosa. Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. *Oklahoma State University*, 942, 2017.

A. Alonso and J. M. Carbó. Machine Learning in Credit Risk: Measuring the Dilemma Between Prediction and Supervisory Cost. *Banco de España*, Working Paper No. 2032, November 2020. ISSN 1579-8666. doi: 10.2139/ssrn.3724374.

E. I. Altman. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, 23(4):589–609, September 1968. doi: 10.1111/j.1540-6261.1968.tb00843.x.

A. F. Atiya. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks*, 12(4):929–935, July 2001. doi: 10.1109/72.935101.

S. Balcaen and H. Ooghe. 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, 38(1):63–93, 2006. ISSN 0890-8389. doi: 10.1016/j.bar.2005.09.001.

BCBS. Principles for the Management of Credit Risk. Bank for International Settlements, Basel, Switzerland, July 1999.

BCBS. Principles for the Management of Credit Risk. Bank for International Settlements, Basel, Switzerland, September 2000.

BCBS. Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version. Bank for International Settlements, Basel, Switzerland, June 2006.

B. Boehmke and B. M. Greenwell. *Hands-On Machine Learning with R*. The R Series. Taylor & Francis Group, Boca Raton, Florida, 1 edition, 2020. ISBN 978-1138495685.

A. Brabazon, J. Dang, I. Dempsey, M. O'Neill, and D. Edelman. Natural Computing in Finance: A Review. *Handbook of Natural Computing*, pages 1707–1735, 2012. doi: 10.1007/978-3-540-92910-9_51.

J. Brownlee. *Machine Learning Mastery with Python: Understand your Data, Create Accurate Models and Work Projects End-To-End*. Machine Learning Mastery. 1.20 edition, August 2021.

T. G. Calderon and J. J. Cheh. A roadmap for future neural research in auditing and risk assessment. *International Journal of Accounting Information Systems*, 3(4):203–236, December 2002. ISSN 1467-0895. doi: 10.1016/S1467-0895(02)00068-4.

C. Chan. What is a ROC Curve and How to Interpret It. `https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/`, September 2020. Accessed: 2021-07-09.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. doi: 10.1613/jair.953.

N. Chen, B. Ribeiro, and A. Chen. Financial credit risk assessment: a recent review. *Artificial Intelligence Review*, 45:1–23, January 2016. doi: 10.1007/s10462-015-9434-x.

A. Chopra and P. Bhilare. Application of Ensemble Models in Credit Scoring Models. *Business Perspectives and Research*, 6(4):129–141, April 2018. doi: 10.1177/2278533718765531.

G. Deng. Analyzing the Risk of Mortgage Default. Master's thesis, UC Berkeley, Fall 2016. URL `https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Grace_Deng_thesis.pdf`.

A. I. Dimitras, S. H. Zanakis, and C. Zopounidis. A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, 90(3):487–513, May 1996. doi: 10.1016/0377-2217(95)00070-4.

G. Donga, K. K. Lai, and J. Yen. Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science*, 1(1):2463–2468, May 2010. doi: 10.1016/j.procs.2010.04.278.

E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi. Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds. January 2021. doi: 10.2139/ssrn.3553781.

EBA. Guidelines on the application of the definition of default under article 178 of Regulation EU No 575/2013. In *CRR Article 178*. Paris, France, September 2016.

S. Finlay. Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210(2):368–378, April 2011. doi: 10.1016/j.ejor.2010.09.029.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, September 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

P. J. FitzPatrick. A comparison of ratios of successful industrial enterprises with those of failed firms. *Certified Public Accountant*, 6:727–731, 1932.

D. Gavalas and T. Syriopoulos. Bank Credit Risk Management and Rating Migration Analysis on the Business Cycle. *International Journal of Financial Studies*, 2(1):122–143, March 2014. doi: 10.3390/ijfs2010122.

T. v. Gerstel and B. Baesens. *Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*. Oxford University Press, Oxford, Jan 2009. ISBN 978–0–19–954511–7. doi: 10.1093/acprof:oso/9780199545117.001.0001.

D. Granström and J. Abrahamsson. Loan default prediction using supervised machine learning algorithms. Master's thesis, KTH, School of Engineering Sciences (SCI), 2019. URL `http://kth.diva-portal.org/smash/get/diva2:1319711/FULLTEXT02.pdf`.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, 46:389–422, January 2002. doi: 10.1023/A:1012487302797.

D. J. Hand and W. E. Henley. Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society Series A*, 160(3):523–541, September 1997. doi: 10.1111/j.1467-985X.1997.00078.x.

K. R. Hasan. Development of a Credit Scoring Model for Retail Loan Granting Financial Institutions from Frontier Markets. *International Journal of Business and Economics Research*, 5:135–142, October 2016. doi: 10.11648/j.ijber.20160505.11.

F. Hu and H. Li. A Novel Boundary Oversampling Algorithm Based on Neighborhood Rough Set Model: NRSBoundary-SMOTE. *Mathematical Problems in Engineering*, 2013:10, November 2013. doi: 10.1155/2013/694809.

S. Hué, C. Hurlin, and S. Tokpavi. Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects. *European Journal of Operational Research*, July 2017. doi: 10.1016/j.ejor.2021.06.053.

J. Jayanthi, K. Joseph, and J. Vaishnavi. Bankruptcy prediction using SVM and Hybrid SVM Survey. *International Journal of Computer Applications*, 33(7):39–45, 2011.

M. Kaya, F. Gurgen, and N. Okay. An Analysis of Support Vector Machines for Credit Risk Modeling. In *Proceedings of the 2008 conference on Applications of Data Mining in E-Business and Finance*, volume 177, pages 25–33, June 2008. doi: 10.3233/978-1-58603-890-8-25.

K. Kennedy, B. M. Namee, and S. J. Delany. Using Semi-Supervised Classifiers for Credit Scoring. *Journal of the Operational Research Society*, 64:513–529, 2013. doi: 10.1057/jors.2011.30.

S. Khemakhem and Y. Boujelbene. Artificial Intelligence for Credit Risk Assessment: Artificial Neural Network and Support Vector Machines. *ACRN Oxford Journal of Finance and Risk Perspectives*, 6(2):1–17, Jan 2017.

M. Kuhn and K. Johnson. *Feature Engineering and Selection: A Practical Approach for Predictive Models*, volume 1st ed. Chapman and Hall/CRC, Boca Raton, FL, 2019. doi: 978-1138079229.

P. R. Kumar and V. Ravi. Bankruptcy prediction in banks and firms via statistical and intelligent techniques. *European Journal of Operational Research*, 180(1):1–28, July 2007. doi: 10.1016/j.ejor.2006.08.043. URL `https://EconPapers.repec.org/RePEc:eee:ejores:v:180:y:2007:i:1:p:1-28`.

M. Lam and B. S. Trinkle. *Using Prediction Intervals to Improve Information Quality of Bankruptcy Prediction Models*, volume 10. Emerald Group Publishing Limited, Bingley, Nov 2014. ISBN 978-1-78441-209-8. doi: 10.1108/S1477-407020140000010014.

J. Le. Decision trees in R. `https://www.datacamp.com/community/tutorials/decision-trees-R`, June 2018. Accessed: 2021-06-14.

S. Lessmann, B. Baesens, H.-V. Seow, and L. Thomas. Subagging for credit scoring models. *European Journal of Operational Research*, May 2015. doi: 10.1016/j.ejor.2015.05.030.

R.-Z. Li, S.-L. Pang, and J.-M. Xu. Neural network credit-risk evaluation model based on back-propagation algorithm. *Proceedings. International Conference on Machine Learning and Cybernetics*, 4:1702–1706, Nov 2002. doi: 10.1109/ICMLC.2002.1175325.

W. Lin, Y.-H. Hu, and C.-F. Tsai. Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42:421–436, 2012. doi: 10.1109/TSMCC.2011.2170420.

A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön. *Machine Learning - A First Course for Engineers and Scientists*. 2021. URL `https://smlbook.org`.

C. X. Ling and V. S. Sheng. *Class Imbalance Problem*. Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning, Boston, MA, 2011. doi: 10.1007/978-0-387-30164-8_110.

R. J. A. Little and D. B. Rubin. Statistical Analysis with Missing. *Journal of Educational Statistics*, 16(2):150–155, Summer 1991. doi: 10.2307/1165119.

H. Liu, H. Motoda, R. Setiono, and Z. Zhao. Feature Selection: An Ever Evolving Frontier in Data Mining. *JMLR: Workshop and Conference Proceedings*, 10:4–13, 2010.

M. Lv, Y. Ren, and Y. Chen. Research on imbalanced data: based on SMOTE-AdaBoost algorithm. In *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, pages 1165–1170, 2019. doi: 10.1109/EITCE47263.2019.9094859.

F. Melo. *Area under the ROC Curve*, pages 38–39. Springer, New York, 2013. ISBN 978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_209.

C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2021. URL `https://christophm.github.io/interpretable-ml-book/`.

S. Moradi and F. M. Rafiei. A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. *Financial Innovation*, 5, March 2019. doi: 10.1186/s40854-019-0121-9.

P. Moura Oliveira, P. Novais, and L. Reis. *Progress in Artificial Intelligence*. Part II: 19th EPIA Conference on Artificial Intelligence, EPIA 2019. January 2019. ISBN 978-3-030-30243-6. doi: 10.1007/978-3-030-30244-3.

H. Nguyen Duc, I. Kamwa, L.-A. Dessaint, and H. Cao-Duc. A novel approach for early detection of impending voltage collapse events based on the support vector machine. *International Transactions on Electrical Energy Systems*, 27, March 2017. doi: 10.1002/etep.2375.

OeNB. Guidelines on Credit Risk Management. [Rating Models and Validation]. Nov 2004. URL `https://www.oenb.at/dam/jcr:1db13877-21a0-40f8-b46c-d8448f162794/rating_models_tcm16-22933.pdf`.

J. A. Ohlson. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1):109–131, Aug 1980. doi: 10.2307/2490395.

G. Paleologo, A. Elisseeff, and G. Antonini. Subagging for credit scoring models. *European Journal of Operational Research*, 201(2):490–499, March 2010. doi: 10.1016/j.ejor.2009.03.008.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

G. Press. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. `https://cutt.ly/cleaning_big_data_time_consuming`, March 2016. Accessed: 2021-06-28.

A. Qader and W. Sihver. Developing an Advanced Internal Ratings-Based Model by Applying Machine Learning. Master's thesis, KTH, School of Engineering Sciences (SCI), 2020. URL `http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-273418`.

S. Raschka and V. Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing, Birmingham, Dec 2019. ISBN 978-1-78995-575-0.

A. K. Reichert, C.-C. Cho, and G. M. Wagner. An Examination of the Conceptual Issues Involved in Developing Credit-Scoring Models. *Journal of Business Economic Statistics*, 1 (2):101–114, 1983. doi: 10.1080/07350015.1983.10509329.

B. Ribeiro, N. Chen, and A. Kovacec. Shaping Graph Pattern Mining for Financial Risk. *Neurocomputing*, 326-327:123–131, January 2019. doi: 10.1016/j.neucom.2017.01.119.

J. Rochet. Capital requirements and the behaviour of commercial banks. *European Economic Review*, 36(5), June 1992. doi: 10.1016/0014-2921(92)90051-W.

D. Roth. Introduction to Machine Learning. In *CS 446 Machine Learning Fall 2016*. Aug 2016. URL `https://www.cis.upenn.edu/~danroth/Teaching/CS446-17/LectureNotesNew/intro/main.pdf`.

M. Saraswat. Practical Guide to Logistic Regression Analysis in R. `https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/logistic-regression-analysis-r/tutorial/`, 2017. Accessed: 2021-06-13.

S. S. Satchidananda and J. B. Simha. Comparing decision trees with logistic regression for credit risk analysis. *The Journal of Finance*, Jan 2006.

D. Sharma. Improving the Art, Craft and Science of Economic Credit Risk Scorecards using Random Forests: Why Credit Scorers and Economists should use Random Forests. June 2011. doi: 10.2139/ssrn.1861535.

T. Shin. Understanding the Confusion Matrix and how to implement it in Python. `https://bit.ly/binary_confusion_matrix`, May 2020. Accessed: 2021-07-07.

S. Shrivastava, P. M. Jeyanthi, and S. Singh. Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. *Cogent Economics Finance*, page 8, 2020. doi: 10.1080/23322039.2020.1729569.

Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, and Y. Wang. Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending. *Information Sciences*, 525, June 2020. doi: 10.1016/j.ins.2020.03.027.

V. Srinivasan and Y. H. Kim. Credit Granting: A Comparative Analysis of Classification Procedures. *The Journal of Finance*, 42(3):665–681, July 1987. doi: 10.2307/2328378.

A. Steenackers and M. Goovaerts. A credit scoring model for personal loans. *Insurance: Mathematics and Economics*, 8(1):31–34, March 1989.

E. M. H. A. M. Tawfik. Machine learning approach for credit score analysis: A case study of predicting mortgage loan defaults. Master's thesis, NOVA Information Management School (NIMS), 2019. URL `http://hdl.handle.net/10362/62427`.

L. C. Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2):149–172, April–June 2000. doi: 10.1016/S0169-2070(00)00034-0.

L. C. Thomas. *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford University Press, April 2009. ISBN 9780199232130. doi: 10.1093/acprof:oso/9780199232130.001.1.

B. Twala. Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, 37(4):3326–3336, April 2010. doi: 10.1016/j.eswa.2009.10.018.

V. Vapnik. The Support Vector Method of Function Estimation. *Nonlinear Modeling*, pages 55–85, 1998. doi: 10.1007/978-1-4615-5703-6_3.

A. Vellido, L. P.J.G., and J. Vaughan. Neural networks in business: A survey of applications (1992-1998). *Expert Systems with Applications*, 17(1):51–70, July 1999. doi: 10.1016/S0957-4174(99)00016-0.

A. Verikas, Z. Kalsyte, M. Bacauskiene, and A. Gelzinis. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. *Soft Computing*, 14:995–1010, Sept 2009. doi: 10.1007/s00500-009-0490-5.

B. K. Wong and Y. Selvi. Neural network applications in finance: A review and analysis of literature (1990-1996). *Information and Management*, 34:129–139, 1998. doi: 10.1016/S0378-7206(98)00050-0.

M. Woźniak, M. Graña, and E. Corchado. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16, April 2013. doi: 10.1016/j.inffus.2013.04.006.

M. B. Yobas, J. N. Crook, and P. Ross. Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, 210(2):368–378, March 2000. doi: 10.1093/imaman/11.2.111.

L. Zhou, S. Pan, J. Wang, and A. Vasilakos. Machine Learning on Big Data: Opportunities and Challenges. *Neurocomputing*, 237:350–361, 2017. doi: 10.1016/j.neucom.2017.01.026.

W. Zhu, N. Zeng, and N. Wang. Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations. *NESUG Proceedings: Health Care and Life Sciences*, January 2010.